


PCT/US 94/10945

BAR CODE LABEL 		U.S. PATENT APPLICATION			
SERIAL NUMBER 08/127,420		FILING DATE 09/27/93	CLASS 435	GROUP ART UNIT 1807	
APPLICANT	RADOJE DRMANAC, WOODRIDGE, IL.				
	CONTINUING DATA*** VERIFIED <div data-bbox="1003 541 1393 667" data-label="Text"> <p>REC'D 24 NOV 1994 WIPO PCT</p> </div>				
	FOREIGN/PCT APPLICATIONS*** VERIFIED <div data-bbox="587 823 1221 961" data-label="Text"> <p>PRIORITY DOCUMENT</p> </div>				
STATE OR COUNTRY IL	SHEETS DRAWING 3	TOTAL CLAIMS 41	INDEPENDENT CLAIMS 7	FILING FEE RECEIVED \$799.00	ATTORNEY DOCKET NO. ARCD:089PAR
ADDRESS	DAVID L. PARKER ARNOLD, WHITE & DURKEE P.O. BOX 4433 HOUSTON, TX 77210				
TITLE	METHODS AND COMPOSITIONS FOR EFFICIENT NUCLEIC ACID SEQUENCING				
This is to certify that annexed hereto is a true copy from the records of the United States Patent and Trademark Office of the application which is identified above. By authority of the COMMISSIONER OF PATENTS AND TRADEMARKS <div data-bbox="462 1465 662 1507" data-label="Text"> <p>OCT 12 1994</p> </div> <div data-bbox="393 1501 435 1516" data-label="Text"> <p>Date</p> </div> <div data-bbox="763 1501 901 1516" data-label="Text"> <p>Certifying Officer</p> </div> <div data-bbox="896 1432 1432 1570" data-label="Text"> <p><i>K. K. Spudgen</i></p> </div>					

BEST AVAILABLE COPY



This Page Blank (uspto)



08/127420 A/Wofec

BACKGROUND OF THE INVENTION

The government may own rights in the present invention pursuant to Department of Energy grant LDRD 03235.

5

1. Field of the Invention

10

The present invention generally relates to the field of molecular biology. The invention particularly provides novel methods and compositions to enable highly efficient sequencing of nucleic acid molecules. The methods of the invention are suitable for sequencing long nucleic acid molecules, including chromosomes and RNA, without cloning or subcloning steps.

15

2. Description of the Related Art

20

Nucleic acid sequencing forms an integral part of scientific research today. Determining the sequence, i.e. the primary structure, of nucleic acid molecules and segments is important in regard to individual research projects investigating a range of particular target areas. Information gained from sequencing impacts science, medicine, agriculture and all areas of biotechnology. Nucleic acid sequencing is, of course, vital to the human genome project and other large-scale undertakings, the aim of which is to further our understanding of evolution and the function of organisms and to provide an insight into the causes of various disease states.

25

30

The Human Genome Project (HGP) is currently underway and sequencing of the entire human genome is in progress at various centres. However, progress in this area is generally both slow and costly. Nucleic acid sequencing is usually determined on polyacrylamide gels which separate DNA fragments in the range of 1 to 500 bp, differing in length by one nucleotide. The actual

determination of the sequence, i.e., the order of the individual A, G, C and T nucleotides may be achieved in two ways. Firstly, using the Maxam and Gilbert method of chemically degrading the DNA fragment at specific nucleotides (Maxam & Gilbert, 1977), or
5 secondly, using the dideoxy chain termination sequencing method described by Sanger and colleagues (Sanger et al., 1977). Both methods are time-consuming and laborious.

10 More recently, other methods of nucleic acid sequencing have been proposed which do not employ an electrophoresis step, these methods may be collectively termed Sequencing By Hybridization or SBH (Drmanac et al., 1991; Cantor et al., 1992; Drmanac & Crkvenjakov, U.S. Patent 5,202,231). Development of certain of
15 these methods has given rise to new solid support type sequencing tools known as sequencing chips. SBH technology has potential for increasing the speed with which nucleic acids can be sequenced, but these methods still suffer from several drawbacks.

20 SBH can be conducted in two basic ways which are often referred to as Format 1 and Format 2 (Cantor et al., 1992). In Format 1, oligonucleotides of unknown sequence, generally of about 100-1000 nucleotides in length, are arrayed on a solid support or filter so that the unknown samples themselves are immobilized (Strezoska et al., 1991; Drmanac & Crkvenjakov, U.S.
25 Patent 5,202,231). Replicas of the array are then interrogated by hybridization with sets of labeled probes of about 6 to 8 residues in length. In Format 2, a sequencing chip is formed from an array of oligonucleotides with known sequences of about 6 to 8 residues in length (Southern, WO 89/10977; Khrapko et al.,
30 1991; Southern et al., 1992). The nucleic acids of unknown sequence are then labeled and allowed to hybridize to the immobilized oligos.

Unfortunately, both of these SBH formats have several limitations, particularly the requirement for prior DNA cloning steps. In Format 1, other significant problems include attaching the various nucleic acid pieces to be sequenced to the solid surface support or preparing a large set of longer probes. In Format 2, major problems include labelling the nucleic acids of unknown sequence, high noise to signal ratios which generally result and the fact that only short sequences can be determined. Therefore, the art would clearly benefit from a new procedure for nucleic acid sequencing, and particularly, one which avoids the tedious processes of cloning and/or subcloning.

SUMMARY OF THE INVENTION

The present invention seeks to overcome these and other drawbacks inherent in the prior art by providing new methods and compositions for the sequencing of nucleic acids. In this novel technique nucleic acid sequences are determined by means of hybridization with two sets of small oligonucleotide probes of known sequences. The methods of the invention allow high discriminatory sequencing of extremely large nucleic acid molecules, including chromosomal material or RNA, without prior cloning, subcloning or amplification. Furthermore, the present methods do not require large numbers of probes, the complex synthesis of longer probes, or the labelling of a complex mixture of nucleic acids segments.

To determine the sequence of a nucleic acid according to the methods of the present invention, one would generally identify sequences from the nucleic acid by sequentially hybridizing with complementary sequences from two sets of small oligonucleotide probes (oligos) of defined length and known sequence, which cover most combinations of sequences for that length of probe. One

would then analyze the sequences identified to determine stretches of the identified sequences which overlap, and reconstruct or assemble the complete nucleic acid sequence from such overlapping sequences.

5

The invention is applicable to sequencing nucleic acid molecules of very long length. As a practical matter, the nucleic acid molecule to be sequenced will generally be fragmented to provide small or intermediate length nucleic acid fragments which may be readily manipulated. The term nucleic acid fragment, as used herein, most generally means a nucleic acid molecule of between about 10 base pairs (bp) and about 100 bp in length. The most preferred methods of the invention are contemplated to be those in which the nucleic acid molecule to be sequenced is treated to provide nucleic acid fragments of intermediate length, i.e., of between about 10 bp and about 40 bp. However, it should be stressed that the present invention is not a method of completely sequencing small nucleic acid fragments, rather it is a method of sequencing nucleic acid molecules *per se*, which involves determining portions of sequence from within the molecule - whether this is done using the whole molecule, or for simplicity, whether this is achieved by first fragmenting the molecule into smaller sized sections of from about 4 to about 1000 bases.

25

Sequences from nucleic acid molecules are determined by hybridizing to small oligonucleotide probes of known sequence. In referring to "small oligonucleotide probes", the term "small" means probes of less than 10 bp in length, and preferably, probes of between about 4 bp and about 9 bp in length. In one exemplary sequencing embodiment, probes of about 6 bp in length are contemplated to be particularly useful. For the sets of oligos to cover all combinations of sequences for the length of probe chosen, their number will be represented by 4^F , wherein F is the

30

length of the probe. For example, for a 4-mer, the set would contain 256 probes; for a 5-mer, the set would contain 1024 probes; for a 6-mer, 4096 probes; a 7-mer, 16384 probes; and the like. The synthesis of oligos of this length is very routine in the art and may be achieved by automated synthesis.

In the methods of the invention, one set of the small oligonucleotide probes of known sequence, which may be termed the first set, will be attached to a solid support, i.e., immobilized on that support in such a way so that they are available to take part in hybridization reactions. The other set of small oligonucleotide probes of known sequence, which may be termed the second set, will be probes which are in solution and which are labelled with a detectable label. The sets of oligos may include probes of the same or different lengths.

The process of sequential hybridization means that nucleic acid molecules, or fragments, of unknown sequence can be hybridized to the distinct sets of oligonucleotide probes of known sequences at separate times (Figure 1). The nucleic acid molecules or fragments will generally be denatured, allowing hybridization, and added to the first, immobilized set of probes under discriminating hybridization conditions to ensure that only fragments with complementary sequences hybridize. Fragments with non-complementary sequences are removed and the next round of discriminating hybridization is then conducted by adding the second, labelled set of probes, in solution, to the combination of fragments and probes already formed. Labelled probes which hybridize adjacent to a fixed probe will remain attached to the support and can be detected, which is not the case when there is space between the fixed and labelled probes (Figure 1).

Nucleic acid sequences which are "complementary" are those which are capable of base-pairing according to the standard

Watson-Crick complementarity rules. That is, that the larger purines will always base pair with the smaller pyrimidines to form only combinations of Guanine paired with Cytosine (G:C) and Adenine paired with either Thymine (A:T), in the case of DNA, or Adenine paired with Uracil (A:U) in the case of RNA.

As used herein, the term "complementary sequences" means nucleic acid sequences which are substantially complementary over their entire length and have very few base mismatches. For example, nucleic acid sequences of six bases in length may be termed complementary when they hybridize at five out of six positions with only a single mismatch. Naturally, nucleic acid sequences which are "completely complementary" will be nucleic acid sequences which are entirely complementary throughout their entire length and have no base mismatches.

After identifying, by hybridization to the oligos of known sequence, various individual sequences which are part of the nucleic acid fragments, these individual sequences are next analyzed to identify stretches of sequences which overlap. For example, portions of sequences in which the 5' end is the same as the 3' end of another sequence, or vice versa, are identified. The complete sequence of the nucleic acid molecule or fragment can then be delineated, i.e., it can be reconstructed from the overlapping sequences thus determined.

The processes of identifying overlapping sequences and reconstructing the complete sequence will generally be achieved by computational analysis. For example, if a labelled probe 5'-TTTTTT-3' hybridizes to the spot containing the fixed probe 5'-AAAAAA-3', a 12-mer sequence of 5'-AAAAATTTTT-3', defined by combining the sequence of the two probes has been determined from the nucleic acid molecule of previously unknown sequence. The next question to be answered is which nucleotide follows next

after the 5'-AAAAAATTTTTT-3' sequence. There are four possibilities represented by the fixed probe 5'-AAATT-3' and labelled probes 5'-TTTTTA-3' for A; 5'-TTTTTT-3' for T; 5'-TTTTTC-3' for C; and 5'-TTTTTG-3' for G. If, for example, 5'-TTTTTC-3' is positive and the other three are negative, then the assembled sequence is extended to 5'-AAAAAATTTTTTC-3'. In the next step, the algorithm determines which of the labelled probes TTTCA, TTCT, TTCC or TTTCG are positive at the spot containing the fixed probe AAATT. The process is repeated until all positive (F + P) oligonucleotide sequences are used or defined as false positives.

The present invention thus provides a very effective way to sequence nucleic acid fragments and molecules of long length. Large nucleic acid molecules, as defined herein, are those molecules which need to be fragmented prior to sequencing. They will generally be of at least about 45 or 50 base pairs (bp) in length, and will most often be longer. In fact, the methods of the invention may be used to sequence nucleic acid molecules with virtually no upper limit on length, so that sequences of about 100 bp, 1 kilobase (kb), 100 kb, 1 megabase (Mb), and 50 Mb or more may be sequenced, up to and including complete chromosomes, such as human chromosomes, which are about 100 Mb in length. Such a large number is well within the scope of the present invention and sequencing this number of bases will require two sets of 8-mers or 9-mers (so that F + P = 16-18). The nucleic acids to be sequenced may be DNA, such as cDNA, genomic DNA, microdissected chromosome bands, cosmid DNA or YAC inserts, or may be RNA, including mRNA, rRNA, tRNA or snRNA.

The process of determining the sequence of a long nucleic acid molecule involves simply identifying sequences of length F + P from the molecule and combining the sequences using a suitable algorithm. In practical terms, one would most likely

first fragment the nucleic acid molecule to be sequenced to produce smaller fragments, such as intermediate length nucleic acid fragments. One would then identify sequences of length $F + P$ by sequentially hybridizing the fragments to complementary sequences from the two sets of small oligonucleotide probes of known sequence, as described above. In this manner, the complete nucleic acid sequence of extremely large molecules can be reconstructed from overlapping sequences of length $F + P$.

Whether the nucleic acid to be sequenced is itself an intermediate length fragment or is first treated to generate such length fragments, the process of identifying sequences from such nucleic acid fragments by hybridizing to two sets of small oligonucleotide probes of known sequence is central to the sequencing methods disclosed herein. This process generally comprises the following steps:

- (a) contacting the set or array of attached or immobilized oligonucleotide probes with the nucleic acid fragments under hybridization conditions effective to allow fragments with a complementary sequence to hybridize sufficiently to a probe, thereby forming primary complexes wherein the fragment has both hybridized and non-hybridized, or "free", sequences;
- (b) contacting the primary complexes with the set of labelled oligonucleotide probes in solution under hybridization conditions effective to allow probes with complementary sequences to hybridize to a non-hybridized or free fragment sequence, thereby forming secondary complexes wherein the fragment is hybridized to both an attached (immobilized) probe and a labelled probe;

(c) removing from the secondary complexes any labelled probes that have not hybridized adjacent to an attached probe, thereby leaving only adjacent secondary complexes;

5

(d) detecting the adjacent secondary complexes by detecting the presence of the label in the labelled probe; and

10

(e) identifying oligonucleotide sequences from the nucleic acid fragments in the adjacent secondary complexes by combining or connecting the known sequences of the hybridized attached and labelled probes.

15

The hybridization or 'washing conditions' chosen to conduct either one, or both, of the hybridization steps may be manipulated according to the particular sequencing embodiment chosen. For example, both of the hybridization conditions may be designed to allow oligonucleotide probes to hybridize to a given nucleic acid fragment when they contain complementary sequences, i.e., substantially matching sequences, such as those sequences which hybridize at five out of six positions. The hybridization steps would preferably be conducted using a simple robotic device as is routinely used in current sequencing procedures.

20

25

Alternatively, the hybridization conditions may be designed to allow only those oligonucleotide probes and fragments which have completely complementary sequences to hybridize. These more discriminating or 'stringent' conditions may be used for both distinct steps of the sequential hybridization process or for either step alone. In such cases, the oligonucleotide probes, whether immobilized or labelled probes, would only be allowed to hybridize to a given nucleic acid fragment when they shared completely complementary sequences with the fragment.

30

5 The hybridization conditions chosen will generally dictate
the degree of complexity required to analyze the data obtained.
Equally, the computer programs available to analyze any data
generated may dictate the hybridization conditions which must be
employed in a given laboratory. For example, in the most
discriminating process, both hybridization steps would be
conducted under conditions that allow only oligos and fragments
with completely complementary sequences to hybridize. As there
will be no mismatched bases, this method involves the least
10 complex computational analyses and, for this reason, it is the
currently preferred method for practicing the invention.
However, the use of less discriminating conditions for one or
both hybridization steps also falls within the scope of the
present invention.

15 Suitable hybridization conditions for use in either or both
steps may be routinely determined by optimization procedures or
'pilot studies'. Various types of pilot studies are routinely
conducted by those skilled in the art of nucleic acid sequencing
20 in establishing working procedures and in adapting a procedure
for use in a given laboratory. For example, conditions such as
the temperature; the concentration of each of the components; the
length of time of the steps; the buffers used and their pH and
ionic strength may be varied and thereby optimized.

25 In preferred embodiments, the nucleic acid sequencing method
of the invention involves a discriminating step to select for
secondary hybridization complexes which include immediately
adjacent immobilized and labelled probes, as distinct from those
30 which are not immediately adjacent and are separated by one, two
or more bases. A variety of processes are available for removing
labelled probes that are not hybridized immediately adjacent to
an attached probe, i.e., not hybridized back to back, each of
which leaves only the immediately adjacent secondary complexes.

Such discriminatory processes may rely solely on washing steps of controlled stringency wherein the hybridization conditions employed are designed so that immediately adjacent probes remain hybridized due to the increased stability afforded by the stacking interactions of the adjacent nucleotides. Again, washing conditions such as temperature, concentration, time, buffers, pH, ionic strength and the like, may be varied to optimize the removal of labelled probes which are not immediately adjacent.

In preferred embodiments the immediately adjacent immobilized and labelled probes would be ligated, i.e., covalently joined, prior to performing washing steps to remove any non-ligated probes. Ligation may be achieved by treating with a solution containing a chemical ligating agent, such as, e.g., water-soluble carbodiimide or cyanogen bromide. More preferably, a ligase enzyme, such as T₄ DNA ligase from T₄ bacteriophage, which is commercially available from many sources (e.g., Biolabs), may be employed. In any event, one would then be able to remove non-immediately adjacent labelled probes by more stringent washing conditions which can not affect covalently connected labeled and fixed probes.

The remaining adjacent secondary complexes would be detected by observing the location of the label from the labelled probes present within the complexes. The oligonucleotide probes may be labeled with a chemically-detectable label, such as fluorescent dyes, or adequately modified to be detected by a chemiluminescent developing procedure, or radioactive labels such as ³⁵S, ³H, ³²P or ³³P. Probes may also be labeled with non-radioactive isotopes and detected by mass spectrometry.

Currently, the most preferred method contemplated for practicing the present invention involves performing the

hybridization steps under conditions designed to allow only those oligonucleotide probes and fragments which have completely complementary sequences to hybridize and which allow only those probes which are immediately adjacent to remain hybridized. This method subsequently requires the least complex computational analysis.

Where the nucleic acid molecule of unknown sequence is longer than about 45 or 50 bp, one effective method for determining its sequence generally involves treating the molecule to generate nucleic acid fragments of intermediate length, and determining sequences from the fragments. The nucleic acid molecule, whether it be DNA or RNA may be fragmented by any one of a variety of methods including, for example, cutting by restriction enzyme digestion, shearing by physical means such as ultrasound treatment, by NaOH treatment or by low pressure shearing.

In certain embodiments, e.g., involving small oligonucleotide probes between about 4 bp and about 9 bp in length, one may aim to produce nucleic acid fragments of between about 10 bp and about 40 bp in length. Naturally, longer length probes would generally be used in conjunction with sequencing longer length nucleic acid fragment, and vice versa. In certain preferred embodiments, the small oligonucleotide probes used will be about 6 bp in length and the nucleic acid fragments to be sequenced will generally be about 20 bp in length. If desired, fragments may be separated by size to obtain those of an appropriate length, e.g., fragments may be run on a gel, such as an agarose gel, and those with approximately the desired length may be excised.

The method for determining the sequence of a nucleic acid molecule may also be exemplified using the following terms.

Initially one would randomly fragment an amount of the nucleic acid to be sequenced to provide a mixture of nucleic acid fragments of length T. One would prepare an array of immobilized oligonucleotide probes of known sequences and length F and a set of labelled oligonucleotide probes in solution of known sequences and length P, wherein $F + P \leq T$ and, preferably, wherein $T \approx 3F$.

One would then contact the array of immobilized oligonucleotide probes with the mixture nucleic acid fragments under hybridization conditions effective to allow the formation of primary complexes with hybridized, complementary sequences of length F and non-hybridized fragment sequences of length $T - F$. Preferably, the hybridized sequences of length F would contain only completely complementary sequences.

The primary complexes would then be contacted with the set of labelled oligonucleotide probes under hybridization conditions effective to allow the formation of secondary complexes with hybridized, complementary sequences of length F and adjacent hybridized, complementary sequences of length P. In preferred embodiments, only those labelled probes with completely complementary sequences would be allowed to hybridize and only those probes which hybridize immediately adjacent to an immobilized probe would be allowed to remain hybridized. In the most preferred embodiments, the adjacent immobilized and labelled oligonucleotide probes would also be ligated at this stage.

Next one would detect the secondary complexes by detecting the presence of the label and identify sequences of length $F + P$ from the nucleic acid fragments in the secondary complexes by combining the known sequences of the hybridized immobilized and labelled probes. Stretches of the sequences of length $F + P$ which overlap would then be identified, thereby allowing the complete nucleic acid sequence of the molecule to be

reconstructed or assembled from the overlapping sequences determined.

5 In the methods of the invention, the oligonucleotides of the first set may be attached to a solid support, i.e. immobilized, by any of the methods known to those of skill in the art. For example, attachment may be via addressable laser-activated photodeprotection (Fodor et al., 1991). A generally preferred method is to attach the oligos through the phosphate group using
10 reagents such as nucleoside phosphoramidite or nucleoside hydrogen phosphate, as described by Southern & Maskos (PCT Patent Application WO 90/03382, incorporated herein by reference), and using glass, nylon or teflon supports.

15 The immobilized oligonucleotides may be formed into an array comprising all probes or subsets of probes of a given length (preferably about 4 to 10 bases), and more preferably, into multiple arrays of immobilized oligonucleotides arranged to form a so-called "sequencing chip". The sequencing chips may be
20 designed for different applications like mapping, partial sequencing, sequencing of targeted regions for diagnostic purposes, mRNA sequencing and large scale genome sequencing. For each application, a specific chip may be designed with different sized probes or with an incomplete set of probes.

25 In one exemplary embodiment, both sets of oligonucleotide probes would be probes of six bases in length, i.e., 6-mers. In this instance, each set of oligos contains 4096 distinct probes. The first set probes is preferably fixed in an array on a
30 microchip, most conveniently arranged in 64 rows and 64 columns. The second set of 4096 oligos would be labeled with a detectable label and dispensed into a set of distinct tubes. In this example, 4096 of the chips would be combined in a large array, or several arrays. After hybridizing the nucleic acid fragments, a

small amount of the labeled oligonucleotides would be added to each microchip for the second hybridization step, only one of each of the 4096 nucleotides would be added to each microchip.

5 Further embodiments of the invention include kits for use in nucleic acid sequencing. Such kits will generally comprise a solid support having attached an array of oligonucleotide probes of known sequences, as shown in Figure 2, wherein the
10 oligonucleotides are capable of taking part in hybridization reactions, and a set of containers comprising solutions of labelled oligonucleotide probes of known sequences.

In the kits, the attached oligonucleotide probes and those
15 in solution may be between about 4 bp and about 9 bp in length, with ones of about 6 bp in length being preferred. The oligos may be labelled with chemically-detectable or radioactive labels, with ³²P-labelled probes being generally preferred. The kits may also comprise a chemical or other ligating agent, such as a DNA
20 ligase enzyme. A variety of other additional compositions and materials may be included in the kits, such as 96-tip or 96-pin devices, buffers, reagents for cutting long nucleic acid molecules and tools for the size selection of DNA fragments. The
25 kits may even include labelled RNA probes so that the probes may be removed by RNAase treatment and the sequencing chips re-used.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1. Basic steps in sequential hybridization process.
30 Step 1: The DNA to be sequenced (T) is hybridized under discriminative conditions to an array of fixed oligonucleotides. Spots with probe Fx and Fy are depicted. Complementary sequences for Fx and Fy are at different positions of T. Step 2: A labeled probe is hybridized to the array. It has a complementary target

on T that is adjacent to the Fx but not to the Fy. Step 3: By applying discriminative conditions or reagents, complexes with no adjacent probes can be selectively melted. Positive signals will be detected only in the case of adjacent probes like Fx and Pi.

5 Fig. 2A. Sequencing chips, representing an array of 4^p identical sections each containing identical (or different) array of oligonucleotides. Sections can be separated by physical barriers or by hydrophobic strips.

10 Fig. 2B is an enlargement of a chip section containing 4^p spots with each with a particular oligonucleotide species synthesized or spotted on that area. Spots can be as small as several microns and the size of the section about 1 mm to 10 mm.

15 Fig. 2C represents a set of tubes (or one or more plates with appropriate number of wells (in this case 4^p wells). Each well contains an amount of a specific labeled oligonucleotide. Additional amounts of the probes can be stored unlabeled if the labeling is not done during synthesis; in this case sequencing
20 kit will contain necessary components for probe labeling. The lines that are connecting tubes/wells with chip sections depict a step in the sequencing procedure where an amount of a labeled probe is transferred to a chip section. The transferring can be
25 done by pipetting (single or multi-channel) or by pin array transferring liquid by surface tension. Transferring tools can be also included in the sequencing kit.

30 Fig. 3. Hybridization of DNA fragments produced by a random cutting of an amount of a DNA molecule. In part 1 DNA fragment T1 is such that contains complete targets for both, fixed and non-fixed-labeled probe. Part 2 represents the case when DNA fragment T is not appropriately cut. In Part 3 there is enough space for probe P to hybridize, but the adjacent sequence is not

complementary to it. Both case 2 and case 3 will reduce signal by saturating even 99% of the molecules of attached probe F. Simultaneous hybridization with DNA fragments and labeled probes and cycling of the hybridization process are some possible ways to increase yield of correct adjacent hybridizations.

DETAILED DESCRIPTION
OF THE PREFERRED EMBODIMENTS

Determining the sequences of nucleic acid molecules is vital to all areas of basic and applied biological research (Drmanac & Crkvenjakov, 1990). The present invention provides new and efficient methods for use in sequencing and analyzing nucleic acid molecules. One intended use for this methodology is, in conjunction with other sequencing techniques, for work on the Human Genome Project (HGP).

Presently, two methods of sequencing by hybridization, SBH, are known. In the first, Format 1, unknown genomic DNAs or oligonucleotides of up to about 100-2000 nucleotides in length are arrayed on a solid substrate. These DNAs are then interrogated by hybridization with a set of labeled probes which are generally 6- to 8-mers. In the inverse technique, Format 2, oligomers of 6 to 8 nucleotides are immobilized on a solid support and allowed to anneal to pieces of cloned and labeled DNA.

In either type of SBH analysis, many steps must be included in order to arrive at a definitive sequence. Particular problems of current SBH methods are those associated with the synthesis of large numbers of probes and the difficulties of effective discriminative hybridization. Full match-mismatch discrimination is difficult due to two main reasons. Firstly, the end mismatch

of probes longer than 10 bases is very indiscriminative, and secondly, the complex mixture of labeled DNA segments which result when analyzing a long DNA fragment generates a high background.

5

The present invention provides effective discriminative hybridization without large numbers of probes or probes of increased length, and also eliminates many of the labeling and cloning steps which are particular disadvantages of each of the known SBH methods. The disclosed highly efficient nucleic acid sequencing methods are based upon hybridization with two sets of small oligonucleotide probes of known sequences, and thus at least double the length of sequence which can be determined. These methods allow extremely large nucleic acid molecules, including chromosomes, to be sequenced and solve various other SBH problems such as, for example, the attachment or labelling of many nucleic acid fragments. The invention is extremely powerful as it may also be used to sequence RNA and even unamplified RNA samples.

20

The nucleic acids to be sequenced may first be fragmented. This may be achieved by any means including, for example, cutting by restriction enzyme digestion; shearing by physical means such as ultrasound treatment; by NaOH treatment, and the like. If desired, fragments of an appropriate length, such as between about 10 bp and about 40 bp may be cut out of a gel. The complete nucleic acid sequence of the original molecule, such as a human chromosome, would be determined by defining F + P sequences present in the original molecule and assembling portions of overlapping F + P sequences. This does not, therefore, require an intermediate step of determining fragment sequences, rather, the sequence of the whole molecule is constructed from F + P sequences delineated.

30

For the purposes of the following discussion, it will be assumed that four bases make up the sequences of the nucleic acids to be sequenced. These are A, G, C and T for DNA and A, G, C and U for RNA. To carry out the invention, one would generally first prepare a number of small oligonucleotide probes of defined length which cover all combinations of sequences for that length of probe. This number is represented by 4^N (4 to the power N) where the length of the probe is termed N . For example, there are 4096 possible sequences for a 6-mer probe ($4^6=4096$).

One set of such probes of length F (4^F) would be fixed in a square array on a microchip - which may be in the range of 1 mm^2 or 1 cm^2 . In the present example, these would be arranged in 64 rows and 64 columns. Naturally, one would ensure that the oligo probes were attached, or otherwise immobilized, to the microchip surface so that were able to take part in hybridization reactions. Another set of oligos of length P , 4^P in number, would be also synthesized. The oligos in this "P set" would be labeled with a detectable label and would be dispensed into a set of tubes (Figure 2).

4^P of the chips would be combined in a large array (or several arrays of approximately $20\text{-}100 \text{ cm}^2$, for a convenient size); where P corresponds to the length of oligonucleotides in the second oligomer set (figure 2). Again, as a convenient example, P is chosen to be six ($P = 6$).

The nucleic acids to be sequenced would be fragmented to give smaller nucleic acid fragments of unknown sequence. The average length of these fragments, termed T , should generally be greater than the combined length of F and P and may be about three times the length of F (i.e., $F + P \leq T$ and $T \sim 3F$). In the present example, one would aim to produce nucleic acid fragments of approximately 20 base pairs in length. These fragments would

be denatured and added to the large arrays under conditions which facilitate hybridization of complementary sequences.

5 In the simplest and currently preferred form of the invention, hybridization conditions would be chosen which would allow significant hybridization to occur only if 6 sequential nucleotides in a nucleic acid fragment were complementary to all 6 nucleotides of an F oligonucleotide probe. Such hybridization conditions would be determined by routine optimization pilot
10 studies in which conditions such as the temperature, the concentration of various components, the length of time of the steps, and the buffers used, including the pH of the buffer.

15 At this stage, each microchip would contain certain hybridized complexes. These would be in the form of probe:fragment complexes in which the entire sequence of the probe is hybridized to the fragment, but in which the fragment, being longer, has some non-hybridized sequences which form a "tail" or "tails" to the complex. In this example, the
20 complementary hybridized sequences would be of length F and the non-hybridized sequences would total T - F in length. The complementary portion of the fragment may be at or towards an appropriate end, so that a single longer non-hybridized tail is formed. Alternatively, the complementary portion of the fragment
25 may be towards the opposite end, so that two non-hybridized tails are formed (Figure 3).

After washing to remove the non-complementary nucleic acid fragments which did not hybridize, a small amount of the labeled
30 oligonucleotides in set P would be added to each microchip for hybridization to the nucleic acid fragment tails of unknown sequence which protrude from the probe:fragment complexes. Only one of each of the 4^p nucleotides would be added to each microchip. Again, it is currently preferred to use hybridization

conditions which would allow significant binding to occur only if all the 6 nucleotides of a labelled probe were complementary to 6 sequential nucleotides of a nucleic acid fragment tail. The hybridization conditions would be determined by pilot studies, as described above, in which components such as the temperature, concentration, time, buffers and the like, were optimized.

At this stage, each microchip would then contain certain 'secondary hybridized complexes'. These would be in the form of probe:fragment:probe complexes in which the entire sequence of each probe is hybridized to the fragment, and in which the fragment likely has some non-hybridized sequences. In these secondary hybridized complexes the immobilized probe and the labelled probe may be hybridized to the fragment so that the two probes are immediately adjacent or "back to back". However, given that the fragments will generally be longer than the sum of the lengths of the probes, the immobilized probe and the labelled probe may be hybridized to the fragment in non-adjacent positions separated by one or more bases.

The large arrays would then be treated by a process to remove the non-hybridized labelled probes. In preferred embodiments, the process employed would remove not only the non-hybridized labelled probes, but also the non-adjacently-hybridized labelled probes from the array. The process would employ discriminating conditions to allow those secondary hybridization complexes which include adjacent immobilized and labelled probes to be discriminating from those secondary hybridization complexes in which the nucleic acid fragment is hybridized to two probes but which probes are not adjacent. This is an important aspect of the invention in that it will allow the ultimate delineation of a section of fragment sequence corresponding to the combined sequences of the immobilized probe and the labelled probe.

The discrimination process employed to remove non-hybridized and non-adjacently-hybridized probes from the array whilst leaving the adjacently-hybridized probes attached may again be a controlled washing process. The adjacently-hybridized probes would be unaffected by the chosen conditions by virtue of their increased stability due to the stacking reactions of the adjacent nucleotides. However, in preferred embodiments, it is contemplated that one would treat the large arrays so that any adjacent probes would be covalently joined, e.g., by treating with a solution containing a chemical ligating agent or, more preferably, a ligase enzyme.

In any event, the complete array would be subjected to stringent washing so that the only label left associated with the array would be in the form of double-stranded probe-fragment-probe complexes with adjacent hybridized portions of length $F + P$ (i.e., 12 nucleotides in the present example). Using this two step hybridization reaction, very high discrimination is possible because three or four independent discriminative processes are taken into account: discriminative hybridization of fragment T to F bases long probe; discriminative hybridization of P bases long probe to fragment T; discriminative stability of full match ($F + T + P$) hybrid in comparison to P hybrids or even to mismatched hybrids containing non-adjacent $F + P$ probes; and discriminative ligation of the two end bases of F and P.

One would then detect the so-called adjacent secondary complexes by observing the location of the remaining label on the array. From the position of the label, $F + P$ (e.g., 12) nucleotide long sequences from the fragment could be determined by combining the known sequences of the immobilized and labelled probes. The complete nucleic acid sequence of the original molecule, such as a human chromosome, could then be reconstructed

or assembled from the overlapping F + P sequences thus determined.

5 When ligation is employed in the sequencing process, as is currently preferred, then the ordinary oligonucleotides chip cannot be reused. The inventor contemplates that this disadvantage may be overcome in various ways. For example, one may generate a specifically cleavable bond between the probes and then cleave the bond after detection. Alternatively, one may
10 employ ribonucleotides for the second probe, probe P, or use a ribonucleotide for the joining base in probe P, so that this probe may subsequently be removed by RNAase or uracil-DNA glycosylase treatment (Craig et al., 1989). Other contemplated methods are to establish bonds by chemical ligation which can be
15 selectively cut (Dolinnaya et al., 1988).

Further variations and improvements to this sequencing methodology are also contemplated and fall within the scope of the present invention. This includes the use of modified
20 oligonucleotides to increase the specificity or efficiency of the methods, similar to that described by Hoheisel & Lehrach (1990). Cycling hybridizations can also be employed to increase the hybridization signal, as is used in PCR technology. In these cases, one would use cycles with different temperatures to re-
25 hybridize certain probes. The invention also provides for determining shifts in reading frames by using equimolar amounts of probes which have a different base at the end position. For example, using equimolar 7-mers in which the first six bases are the same defined sequence and the last position may be A, T, C or
30 G in the alternative.

The following examples are included to demonstrate preferred embodiments of the invention. It should be appreciated by those

of skill in the art that the techniques disclosed in the examples which follow represent techniques discovered by the inventor to function well in the practice of the invention, and thus can be considered to constitute preferred modes for its practice.

5 However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the specific embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the invention.

10

EXAMPLE I

PREPARATION OF SUPPORT BOUND OLIGONUCLEOTIDES

15 Support bound oligonucleotides may be prepared by any of the methods known to those of skill in the art using any suitable support such as glass, polystyrene or teflon. One strategy is to precisely spot oligonucleotides synthesized by standard synthesizers, and another strategy uses the strong biotin-streptavidin interaction as a linker.

20

Modified oligonucleotides may be used throughout the procedures of the present invention to increase the specificity or efficiency. A way to achieve this is the substitution of
25 natural nucleotides by base modification, like pyrimidines with a halogen at the C-position, which reportedly influences the base stacking and 2,6-diaminopurine with a third hydrogen bond in its base pairing with thymine which is known to thermally stabilize DNA-duplexes, as disclosed by Hoheisel & Lehrach (1990,
30 incorporated herein by reference). The use of cationic detergents in renaturation of complementary DNA strands is also contemplated for use in accordance herewith, as described by Pontius & Berg (1991, incorporated herein by reference).

It is contemplated that a preferred method will be that described in PCT Patent Application WO 90/03382 (Southern & Maskos), incorporated herein by reference. This method of preparing an oligonucleotide bound to a support involves
5 attaching a nucleoside 3'-reagent through the phosphate group by a covalent phosphodiester link to aliphatic hydroxyl groups carried by the support. The oligonucleotide is then synthesized on the supported nucleoside and protecting groups removed from the synthetic oligonucleotide chain under standard conditions
10 which do not cleave the oligonucleotide from the support. Suitable reagents include nucleoside phosphoramidite and nucleoside hydrogen phosphorate.

An on-chip strategy for the preparation of DNA probe arrays
15 may be employed. For example, addressable laser-activated photodeprotection may be employed in the chemical synthesis of oligonucleotides directly on a glass surface, as described by Fodor et al. (1991), incorporated herein by reference. Probes may also be immobilized on nylon supports as described by Van
20 Ness et al. (1991); or linked to teflon using the method of Duncan & Cavalier (1988), all references being specifically incorporated herein by reference.

25 **EXAMPLE II**

PREPARATION OF SEQUENCING CHIPS

The present example describes physical embodiments of sequencing chips contemplated by the inventor.
30

A basic example is using 6-mers attached to 50 micron surfaces to give a chip with dimensions of 3 x 3 mm which can be combined to give an array of 20 x 20 cm.

Another example is using 9-mer oligonucleotides attached to 10 x 10 microns surface to create a 9-mer chip, with dimensions of 5 x 5 mm. 4000 units of this array will create a 30 x 30 cm chip. Such arrays may be separated physically from each other or by hydrophobic surfaces.

In one example, 4000 labeled 6-mers in 42 96-well plates would be stored. In this case, $F = 9$; $P = 6$; and $F + P = 15$. Chips may have probes of formula B_xN_y , where x is a number of specified bases B and y is a number of non-specified bases, so that $y = 1$ to 4 and $x = 4$ to 10. To achieve more efficient hybridization and to avoid potential influence of the support oligonucleotide, the unspecified bases can be surrounded by specified bases, thus represented by a formula such as NzB_xN_y .

EXAMPLE III

PREPARATION OF NUCLEIC ACID FRAGMENTS AND LABELLED PROBES

The nucleic acids to be sequenced may be obtained from any appropriate source, such as cDNAs, genomic DNA, chromosomal DNA, microdissected chromosome bands, cosmid or YAC inserts, and RNA, including mRNA without any amplification steps. The nucleic acids would then be fragmented by any of the methods known to those of skill in the art including, for example, using restriction enzymes, shearing by ultrasound, NaOH treatment and low pressure shearing (Schriefer et al., 1990; incorporated herein by reference).

The oligonucleotide probes may be prepared by automated synthesis, which is routine to those of skill in the art (Sambrook et al., 1989). Alternatively, probes may be prepared using Genosys Biotechnologies Inc. methods using stacks of porous Teflon wafers.

Oligonucleotide probes may be labelled with, for example, radioactive labels (^{35}S , ^{32}P , ^{33}P) for arrays with 100-200 μm spots; non-radioactive isotopes (Jacobsen et al., 1990); or fluorophores (Brumbaugh et al., 1988). All such labelling methods are routine in the art, as exemplified by further references such as Schubert et al. (1990), Murakami et al. (1991) and Cate et al. (1991), all references being specifically incorporated herein by reference.

EXAMPLE IV

CONDUCTING SEQUENCING BY TWO STEP HYBRIDIZATION

Following are certain examples to describe the execution of the sequencing methodology contemplated by the inventor.

First, the whole chip would be hybridized with mixture of DNA as complex as 100 million of bp (one human chromosome). After proper washing using a simple robotic device on each 5 x 5mm array, one labeled 6-mer would be added. A 96-tip or 96-pin device would be used, performing this in 42 operations.

When using a ligation process, the enzyme could be added with the labeled probes or after the proper washing step to reduce the background. After final washing appropriate for discriminating detection of hybridized adjacent oligonucleotides of length (F + P), as discussed above, signals are scored per each of billion points. It would not be necessary to hybridize all 4000 5 x 5mm arrays at a time and the successive use of smaller number of arrays is possible.

Cycling hybridizations are one possible method for increasing the hybridization signal. In one cycle, most of the fixed probes will hybridize with DNA fragments with tail

sequences non-complementary for labelled probes. By increasing the temperature, those hybrids will be melted. In the next cycle, some of them (~0.1%) will hybridize with an appropriate DNA fragment and additional labeled probes will be ligated. In this case, there occurs a discriminative washing of DNA hybrids for both probe sets simultaneously.

The procedure described herein allows complex chip manufacturing using standard synthesis and precise spotting of oligonucleotides because a relatively small number of oligonucleotides are necessary. For example if all 7-mer oligos are synthesized (16384 probes), lists of 256 million 14-mers can be determined.

One important variant of the invented method is to use more than one differently labeled probe per basic array. This can be executed with two purposes in mind; multiplexing to reduce number of separately hybridized arrays; or to determine a list of even longer oligosequences such as 3 x 6 or 3 x 7. In this case if two labels are used the specificity of the 3 consecutive oligonucleotides can be almost absolute because positive sites must have enough signals of both labels.

A further and additional variant is to use chips containing BxNy probes with y being from 1 to 4. Those chips allow sequence reading in different frames. This can also be achieved by using appropriate sets of labeled probes or both F and P probes could have some unspecified end positions (i.e., some element of terminal degeneracy).

EXAMPLE V

RE-USING SEQUENCING CHIPS

When ligation is employed in the sequencing process, then the ordinary oligonucleotides chip cannot be reused. The inventor contemplates that this disadvantage may be overcome in various ways.

Firstly, one may employ ribonucleotides for the second probe, probe P, so that this probe may subsequently be removed by RNAase treatment. One may also specifically use the uracil base, as described by Craig et al. (1989), incorporated herein by reference.

Secondly, one could generate a specifically cleavable bond between the probes and then cleave the bond after detection. For example, this may be achieved by chemical ligation as described by Shabarova et al. (1991) and Dolinnaya et al. (1988), both references being specifically incorporated herein by reference.

EXAMPLE VI

ANALYZING THE DATA OBTAINED

From the position of the label detected, F + P nucleotide sequences from the fragments would be determined by combining the known sequences of the immobilized and labelled probes corresponding to the labelled positions. The complete nucleic acid sequence of the original molecule, such as a human chromosome, would then be assembled from the overlapping F + P sequences determined by computational deduction.

The processes of computational deduction would employ computer programs using existing algorithms (see, e.g., Drmanac

et al., 1991). If, in addition to $F + P$, $F(\text{space } 1)P$,
 $F(\text{space } 2)P$, $F(\text{space } 3)P$ or $F(\text{space } 4)P$ are determined,
algorithms will be used to match all data sets to correct
potential errors or to solve the situation where there is a
5 branching problem (see, e.g., Drmanac et al., 1989; Bains et al.,
1988)

* * *

10

While the compositions and methods of this invention have
been described in terms of preferred embodiments, it will be
apparent to those of skill in the art that variations may be
15 applied to the composition, methods and in the steps or in the
sequence of steps of the method described herein without
departing from the concept, spirit and scope of the invention.
More specifically, it will be apparent that certain agents which
are both chemically and physiologically related may be
20 substituted for the agents described herein while the same or
similar results would be achieved. All such similar substitutes
and modifications apparent to those skilled in the art are deemed
to be within the spirit, scope and concept of the invention as
defined by the appended claims. All claimed matter and methods
25 can be made and executed without undue experimentation.

REFERENCES

The following references, to the extent that they provide exemplary procedural or other details supplementary to those set forth herein, are specifically incorporated herein by reference.

Bains et al., 1988, A Novel Method for Nucleic Acid Sequence Determination, J. Theor. Biol., 135:303-307.

Brumbaugh et al., 1988, Continuous, on-line DNA sequencing using oligodeoxynucleotide primers with multiple fluorophores, Proc. Natl. Acad. Sci. U.S.A., 85:5610-5614.

Cantor et al., 1992, Genomics, 13, 1378

Cate et al., 1991, Genomic Southern Analysis with Alkaline-Phosphatase-Conjugated Oligonucleotide Probes and the Chemiluminescent Substrate AMPPD, GATA, 8(3):102-106.

Craig et al., 1989, Labelling oligonucleotides to high specific activity, Nucleic Acids Research, 17(12):4605.

Dolinnaya et al., 1988, Site-directed modification of DNA duplexes by chemical ligation, Nucleic Acids Research, 16(9):3721-3738.

Drmanac et al., 1989, Sequencing of megabase plus DNA by hybridization: Theory of the method, Genomics, 4:114-128

Drmanac et al., 1991, An algorithm for the DNA sequence generation from k-tuple word contents of the minimal number of random fragments, J. Biomol. Struct. & Dyn., 8:1085.

Drmanac & Crkvenjakov, 1990, *Scientia Yugoslavica*, 16, 97

Drmanac & Crkvenjakov, U.S. Patent 5,202,231

5 Drmanac et al., 1991, In "Electrophoreses, Supercomputers and the Human Genome", pp 47-59, World Scientific Publishing Co., Singapore.

10 Duncan & Cavalier, 1988, Affinity Chromatography of a Sequence-Specific DNA Binding Protein Using Teflon-Linked Oligonucleotides, *Analytical Biochemistry*, 169:104-108.

15 Fodor et al., 1991, Light-Directed, Spatially Addressable Parallel Chemical Synthesis, *Research Article*, 251:767-768.

20 Hoheisel & Lehrach, 1990, Quantitative measurements on the duplex stability of 2,6-diaminopurine and 5-chloro-uracil nucleotides using enzymatically synthesized, 274(1,2):103-106.

25 Jacobsen et al., 1990, An Approach to the Use of Stable Isotopes for DNA Sequencing, *Genomics*, 8:001-007.

30 Khrapko et al., 1991, *J. DNA Sequencing Mapping*, 1, 375

35 Maxam & Gilbert, 1977, *Proc. Natl. Acad. Sci.*, 74, 560

40 Murakami et al., 1991, Fluorescent-labeled oligonucleotide probes: detection of hybrid formation in solution by fluorescence polarization spectroscopy, *Nucleic Acids Research*, 19(15):4097-4102.

Pontius & Berg, 1991, Rapid renaturation of complementary DNA strands mediated by cationic detergents: A role of high-

probability binding domains in enhancing the kinetics of molecular assemble processes, Proc. Natl. Acad. Sci. U.S.A., 88:8237-8241.

- 5 Rasmussen et al., 1991, Covalent Immobilization of DNA onto Polystyrene Microwells: The Molecules Are Only Bound at the 5' End, Analytical Biochemistry, 198:138-142.
- 10 Sambrook et al., 1989, Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory. Cold Spring Harbor, NY.
- Sanger, et al., 1977, Proc. Natl. Acad. Sci., 74, 5463
- 15 Schriefer et al., 1990, Low pressure DNA shearing: A method for random DNA sequence analysis, Nucleic Acids Research, 18(24):7455.
- 20 Schubert et al., 1990, One-step labeling of oligonucleotides with fluorescein during automated synthesis, Nucleic Acids Research, 18(11):3427.
- 25 Shabarova et al., 1991, Chemical ligation of DNA: the first non-enzymatic assembly of a biologically active gene. Nucleic Acids Research, 19(15):4247-4251.
- Southern, PCT Patent Application WO 89/10977
- Southern & Maskos, PCT Patent Application WO 90/03382
- 30 Southern et al., 1992, Genomics, 13, 1008
- Strezoska et al., 1991, Proc. Natl. Acad. Sci., 86, 10089

Van Ness et al., 1991, A versatile solid support system for
oligodeoxynucleotide probe-based hybridization assays,
Nucleic Acids Research, 19(12):3345.

WHAT IS CLAIMED IS:

1. A method for determining the sequence of a nucleic acid molecule, comprising the steps of:

- (a) identifying sequences from the molecule by sequentially hybridizing the molecule to complementary sequences from two sets of small oligonucleotide probes of known sequence, wherein the first set of probes are attached to a solid support and the second set of probes are labelled probes in solution;
- (b) identifying overlapping stretches of sequence from the sequences identified in step (a); and
- (c) assembling the nucleic acid sequence of the molecule from said overlapping sequences identified.

2. A method for determining the sequence of a nucleic acid molecule, comprising the steps of:

- (a) fragmenting the nucleic acid molecule to be sequenced to provide intermediate length nucleic acid fragments;
- (b) identifying sequences from said fragments by sequentially hybridizing the fragments to complementary sequences from two sets of small oligonucleotide probes of known sequence, wherein the first set of probes are attached to a solid support and the second set of probes are labelled probes in solution;

(c) identifying overlapping stretches of sequence from said sequences identified in step (b); and

5

(d) assembling the nucleic acid sequence of the molecule from said overlapping sequences identified.

10 3. The method of claim 2, wherein said intermediate length nucleic acid fragments are between about 10 bp and about 40 bp in length and said small oligonucleotide probes are between about 4 bp and about 9 bp in length.

15 4. The method of claim 2, wherein said oligonucleotide probes hybridize to completely complementary sequences from said fragments.

20 5. The method of claim 2, wherein said oligonucleotide probes hybridize to immediately adjacent sequences from said fragments.

25 6. The method of claim 5, wherein said oligonucleotide probes hybridize to completely complementary and immediately adjacent sequences from said fragments.

30 7. The method of claim 5, wherein said immediately adjacent oligonucleotide probes are subsequently ligated.

8. The method of claim 1, wherein the hybridization is carried out in cycles.

9. The method of claim 2, wherein step (b) comprises the steps of:

- 5 (a) contacting said first set of small attached oligonucleotide probes with said intermediate length nucleic acid fragments under hybridization conditions effective to allow only those fragments with a completely complementary sequence to hybridize to a probe, thereby forming primary complexes wherein the
10 fragment has hybridized and free sequences;
- 15 (b) contacting said primary complexes with said second set of small labelled oligonucleotide probes under hybridization conditions effective to allow only those probes with completely complementary sequences to hybridize to a free fragment sequence, thereby forming secondary complexes wherein the fragment is hybridized to an attached probe and a labelled probe;
- 20 (c) removing from said secondary complexes labelled probes that are not immediately adjacent to an attached probe, thereby leaving only adjacent secondary complexes;
- 25 (d) detecting said adjacent secondary complexes by detecting the presence of the label; and
- 30 (e) identifying sequences from the nucleic acid fragments in said adjacent secondary complexes by connecting the known sequences of the hybridized attached and labelled probes.

10. A method of nucleic acid sequencing comprising the steps of:

- 5
- (a) fragmenting the nucleic acid to be sequenced to provide nucleic acid fragments of length T;
- 10
- (b) preparing an array of immobilized oligonucleotide probes of known sequences and length F and a set of labelled oligonucleotide probes in solution of known sequences and length P, wherein $F + P \leq T$;
- 15
- (c) contacting said array of immobilized oligonucleotide probes with said nucleic acid fragments under hybridization conditions effective to allow the formation of primary complexes with hybridized, completely complementary sequences of length F and non-hybridized fragment sequences of length $T - F$;
- 20
- (d) contacting said complexes with said set of labelled oligonucleotide probes under hybridization conditions effective to allow only the formation of secondary complexes with hybridized, completely complementary sequences of length F and immediately adjacent hybridized, completely complementary sequences of length P;
- 25
- (e) detecting said secondary complexes by detecting the presence of the label;
- 30
- (f) identifying sequences of length $F + P$ from the nucleic acid fragments in said secondary complexes by combining the known sequences of the hybridized immobilized and labelled probes;

(g) determining stretches of said sequences of length $F + P$ which overlap; and

(h) assembling the complete nucleic acid sequence from said overlapping sequences.

11. The method of claim 10, wherein length T is about three times longer than length F.

12. The method of claim 10, wherein length T is between about 10 bp and about 40 bp, length F is between about 4 bp and about 9 bp and length P is between about 4 bp and about 9 bp.

13. The method of claim 12, wherein length T is about 20 bp, length F is about 6 bp and length P is between about 6 bp.

14. The method of claim 10, wherein said immediately adjacent immobilized and labeled oligonucleotide probes are ligated.

15. A method of nucleic acid sequencing comprising the steps of:

(a) fragmenting the nucleic acid to be sequenced to provide intermediate length nucleic acid fragments;

(b) contacting an array of immobilized small oligonucleotide probes of known sequences with said nucleic acid fragments under hybridization conditions effective to allow only those fragments with a completely complementary sequence to hybridize to a

probe, thereby forming primary complexes wherein the fragment has hybridized and non-hybridized sequences;

- 5 (c) contacting said primary complexes with a set of
labelled small oligonucleotide probes in solution of
known sequences under hybridization conditions
effective to allow only those probes with completely
complementary sequences to hybridize to a non-
10 hybridized fragment sequence, thereby forming secondary
complexes wherein the fragment is hybridized to an
immobilized probe and a labelled probe;
- 15 (d) removing from said secondary complexes labelled probes
that are not immediately adjacent to an immobilized
probe, thereby leaving only adjacent secondary
complexes;
- 20 (e) detecting said adjacent secondary complexes by
detecting the presence of the label;
- 25 (f) identifying sequences from the nucleic acid fragments
in said adjacent secondary complexes by combining the
known sequences of the hybridized immobilized and
labelled probes;
- 30 (g) determining stretches of said sequences which overlap;
and
- (h) assembling the complete nucleic acid sequence from said
overlapping sequences identified.

16. The method of claim 15, wherein the nucleic acid is cloned
DNA or chromosomal DNA.

17. The method of claim 15, wherein the nucleic acid is mRNA.

18. The method of claim 15, wherein the nucleic acid is
5 fragmented by restriction enzyme digestion, ultrasound treatment,
NaOH treatment or low pressure shearing.

19. The method of claim 15, wherein the nucleic acid fragments
10 are between about 10 bp and about 100 bp in length.

20. The method of claim 15, wherein the oligonucleotide probes
15 are between about 4 bp and about 9 bp in length.

21. The method of claim 20, wherein the oligonucleotide probes
are about 6 bp in length.

22. The method of claim 15, wherein said immobilized
20 oligonucleotides are attached to a glass, polystyrene or teflon
solid support.

23. The method of claim 15, wherein said immobilized
25 oligonucleotides are attached to a solid support via a
phosphodiester linkage.

24. The method of claim 15, wherein the labelled oligonucleotide
30 probes are labelled with a non-radioactive isotope or a
chemiluminescent dye.

25. The method of claim 15, wherein the labelled oligonucleotide probes are labelled with ³⁵S, ³²P or ³³P.

5 26. The method of claim 15, wherein labelled probes which are not immediately adjacent to an immobilized probe are removed from the secondary complexes by stringent washing conditions.

10 27. The method of claim 15, wherein labelled probes which are immediately adjacent to an immobilized probe are ligated to said immobilized probe and non-ligated labelled probes are subsequently removed by washing.

15 28. The method of claim 27, wherein said adjacent probes are ligated enzymatically.

20 29. The method of claim 15, wherein multiple arrays of immobilized oligonucleotides are arranged in the form of a sequencing chip.

25 30. A method of nucleic acid sequencing comprising the steps of:

(a) fragmenting the nucleic acid to be sequenced to provide nucleic acid fragments of between about 10 bp and about 40 bp in length;

30 (b) contacting an array of immobilized oligonucleotide probes with known sequences of between about 4 bp and about 9 bp in length with said nucleic acid fragments under hybridization conditions effective to allow only

those fragments with a completely complementary sequence to hybridize to a probe, thereby forming primary complexes wherein the fragment has hybridized and non-hybridized sequences;

- 5
- (c) contacting said complexes with a set of ^{32}P -labelled or ^{33}P -labelled oligonucleotide probes with known sequences of between about 4 bp and about 9 bp in length under hybridization conditions effective to
- 10 allow only those labelled probes with completely complementary sequences to hybridize to a non-hybridized fragment sequence, thereby forming secondary complexes wherein the fragment is hybridized to an immobilized probe and a ^{32}P -labelled or ^{33}P -labelled probe;
- 15
- (d) ligating the immobilized probes and labelled probes which are immediately adjacent with a DNA ligase enzyme, thereby forming ligated secondary complexes;
- 20
- (e) removing from the secondary complexes any non-ligated labelled probes;
- (f) detecting said ligated secondary complexes by detecting the presence of the ^{32}P or ^{33}P label;
- 25
- (g) identifying sequences from the nucleic acid fragments in said ligated secondary complexes by combining the known sequences of the ligated probes;
- 30
- (h) determining stretches of said sequences which overlap; and -

- (i) assembling the complete nucleic acid sequence from said overlapping sequences.

5 31. A kit for use in nucleic acid sequencing, comprising a solid support having attached an array of oligonucleotide probes of known sequences, said oligonucleotides being capable of taking part in hybridization reactions, and a set of containers comprising solutions of labelled oligonucleotide probes of known sequences.

10 32. The kit of claim 31, wherein multiple arrays of immobilized oligonucleotide probes are arranged in the form of a sequencing chip.

15 33. The kit of claim 31, wherein the oligonucleotide probes are between about 4 bp and about 9 bp in length.

20 34. The kit of claim 33, wherein the oligonucleotide probes are about 6 bp in length.

25 35. The kit of claim 31, wherein the oligonucleotide probes are attached to a glass, polystyrene or teflon solid support.

30 36. The kit of claim 31, wherein the oligonucleotide probes are attached to a solid support via a phosphodiester linkage.

37. The kit of claim 31, wherein the labelled oligonucleotide probes are labelled with a non-radioactive isotope or a chemiluminescent dye.

5

38. The kit of claim 31, wherein the labelled oligonucleotide probes are labelled with ^{35}S , ^{32}P or ^{33}P .

10

39. The kit of claim 31, further comprising a ligating agent.

40. The kit of claim 39, wherein the ligating agent is a DNA ligase enzyme.

15

41. An DNA molecule having the structure BxNy wherein y is from 1 to 4.



Disclosed are novel methods and compositions for rapid and highly efficient nucleic acid sequencing based upon hybridization with two sets of small oligonucleotide probes of known sequences. Extremely large nucleic acid molecules, including chromosomes and non-amplified RNA, may be sequenced without prior cloning or subcloning steps. The methods of the invention also solve various current problems associated with sequencing technology such as, for example, high noise to signal ratios and difficult discrimination, attaching many nucleic acid fragments to a surface, preparing many, longer or more complex probes and labelling more species.

15

20

25

g:\arcd\089\pa\01.fus



PATENT
ARCD:089

DECLARATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled "METHODS AND COMPOSITIONS FOR EFFICIENT NUCLEIC ACID SEQUENCING," the Specification of which:

_____ is attached hereto.
X was filed on September 27, 1993 as Application Serial No. 08/127,420.

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims.

I acknowledge the duty to disclose to the Patent and Trademark Office all information known to me to be material to patentability of the subject matter claimed in this application, as "materiality" is defined in Title 37, Code of Federal Regulations, § 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, § 119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

PRIOR FOREIGN APPLICATION(S)

Priority
Claimed

(Number) (Country) (Date Filed)

Yes/No

(Number) (Country) (Date Filed)

Yes/No

I hereby claim the benefit under Title 35, United States Code, § 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, § 112, I acknowledge the duty to disclose all information known to me to be material to patentability of the subject matter claimed in this application, as "materiality" is defined in Title 37, Code of Federal Regulations, § 1.56, which become available between the filing date of the prior application and the national or PCT international filing date of this application:

(Application Serial No.) (Filing Date) (Status)

(Application Serial No.) (Filing Date) (Status)

I hereby direct that all correspondence and telephone calls be addressed to David L. Parker, Arnold, White & Durkee, P.O. Box 4433, Houston, Texas 77210 (512) 320-7200.

I hereby declare that all statements made of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful

false statements may jeopardize the validity of the application or any patent issued thereon.

1-00

Inventor's Full Name: Radoje Drmanac
First Middle Last

Inventor's Signature: *Radoje Drmanac*

Date: 11/9/1993 Country of Citizenship: YUGOSLAVIA

Residence Address : 2622 BURR RIDGE CT. APT # 212
(Include number, street name, city, state, and country)
WOODBRIDGE, IL 60517

Post Office Address:
(if different from
residence address)

G:\arcd\089\pto\declar



PATENT

IN-THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	\$	
Radoje Drmanac	\$	
Serial No.:	\$	Examiner: Unknown
08/127,420	\$	
Filed: September 27, 1993	\$	Group Art Unit: Unknown
	\$	
For: METHODS AND COMPOSITIONS	\$	Atty. Dkt.: ARCD:089/PAR
FOR EFFICIENT NUCLEIC	\$	
ACID SEQUENCING	\$	

DECLARATION CLAIMING SMALL ENTITY STATUS
37 C.F.R. §§ 1.9(f) and 1.27(d) - NONPROFIT ORGANIZATION

Commissioner of Patents
and Trademarks
Washington, D.C. 20231

Sir:

I hereby declare that I am an official empowered to act on behalf of the nonprofit organization identified below:

Name of Organization: ARCH Development Corp.

Address of Organization: 1101 E. 58th Street
The University of Chicago
Chicago, Illinois 60637

The type of organization is a university.

I hereby declare that the organization identified above qualifies as a nonprofit organization as defined in 37 C.F.R. § 1.9(e) (1), and thus is a "small entity" as defined in § 1.9(f), for purposes of paying reduced fees under Sections 41(a) and (b) of Title 35, United States Code, with regard to the above-referenced application.

I hereby declare that exclusive rights to the invention have been conveyed to and remain with the organization, with respect to the above-referenced invention, nor have I assigned, granted, conveyed or licensed and am under no obligation under contract or law to assign, grant, convey or license, any rights in the invention to any person who could not be classified as an independent inventor under 37 CFR § 1.9(c) if that person had made the invention, or to any concern which would not qualify as a small business concern under 37 CFR § 1.9(d) or a nonprofit organization under 37 CFR § 1.9(e), with the exception that the Government may have rights in the invention pursuant to a funding agreement under 35 U.S.C. § 202(c) (4): Department of Energy Grant No. LDRD 03235 and W-31-109-ENG-38.

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is no longer appropriate.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief

are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application, any patent issuing thereon, or any patent to which this verified statement is directed.

ARCH Development Corp.

By: 

Name: Steven Lazarus

Title: President and CEO

Date: 4/16/93

g:\arcd\089\pto\sentry

PATENT

Please direct all communications as follows:

David L. Parker, Esq.
ARNOLD WHITE & DUNKEE
P.O. Box 4433
Houston, Texas 77210
(512) 320-7200

ASSIGNEE:

ARCH Development Corp.

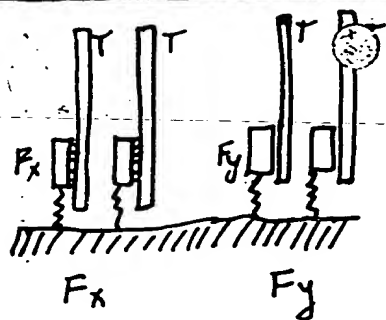
By: *SL*

Name: Steven Lazarus
Title: President and CEO

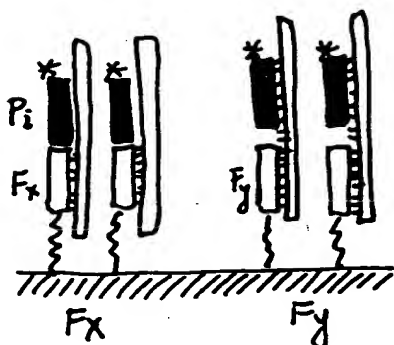
Date: 11/16/93

ASSIGNMENT: Concurrently Filed

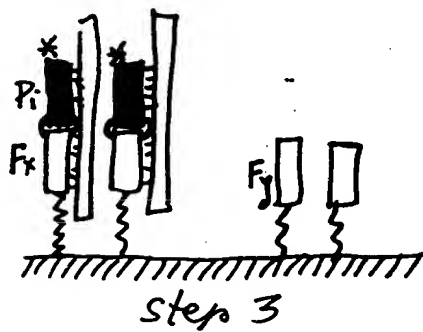
G:\arcd\089\pto\poa



step 1



step 2



step 3

Fig. 1

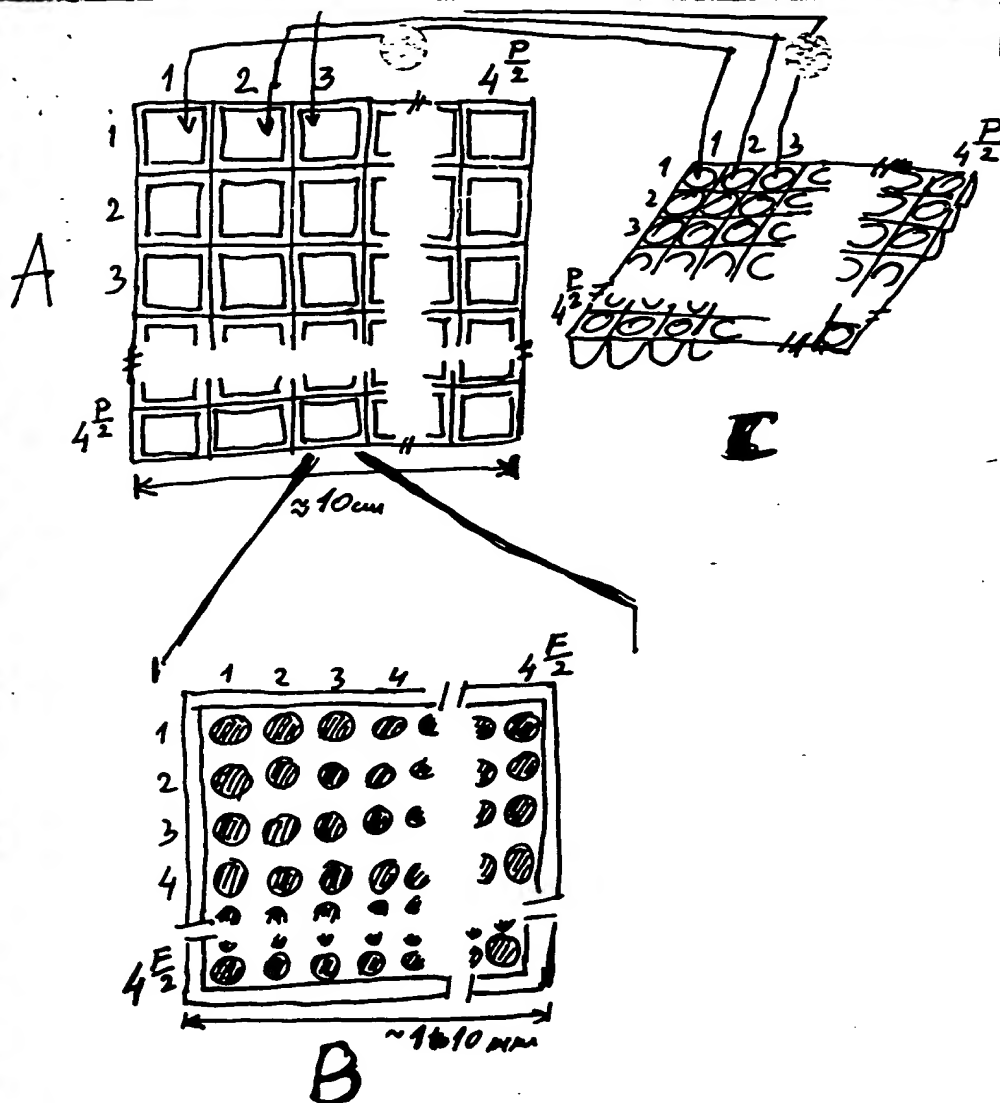


Fig. 2.

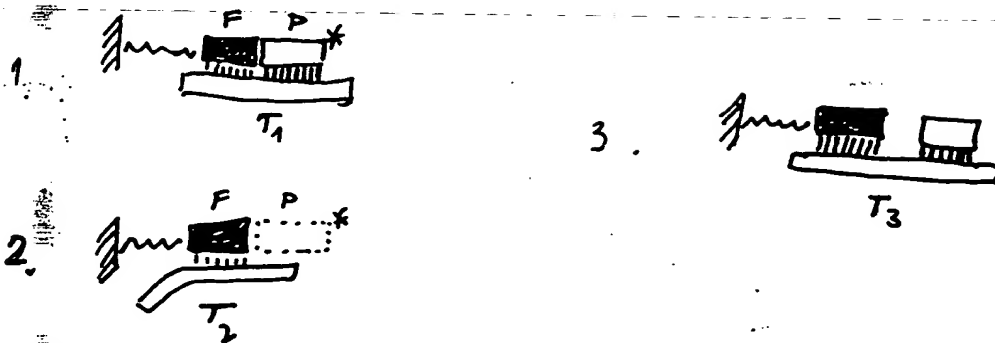


Fig. 3

This Page Blank (uspto)

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

☒ **BLACK BORDERS**

☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**

☒ **FADED TEXT OR DRAWING**

☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**

☐ **SKEWED/SLANTED IMAGES**

☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**

☐ **GRAY SCALE DOCUMENTS**

☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**

☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**

☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

This Page Blank (uspto)